MANUSCRIPT
WRE # 364

Variance of Load Estimates Derived by Piece-Wise Interpolation

Accepted for publication in the
Journal of Environmental Engineering
For November, 1998
Paper no.: 16817

August 1998

By

George Shih
Water Resources Evaluation Department
Xiaosong Wang
Regulation Department
H.J. Grimshaw
Everglades Restoration Department
Joel VanArman
Planning Department

SOUTH FLORIDA WATER MANAGEMENT DISTRICT
3301 Gun Club Road
West Palm Beach, Florida 33406

(Note: Minor revisions were made by the editing staff of the Journal to conform to their format. For exact copies, please contact:
ASCE Reprints,
1801 Alexander Bell Drive,
Reston, VA 20191-4400.)

Variance of Load Estimates Derived by Piece-wise Linear Interpolation

George Shih[1], Xiaosong Wang,[2] H.J. Grimshaw,[3] and Joel VanArman[4]

Abstract: Piece-wise linear interpolation (PLI) is frequently used in environmental studies to estimate missing data. However, to evaluate the reliability of these estimates, the variances of these interpolated values must be quantified.

We propose a procedure to quantify this PLI variance which involves establishing a semi-variogram with coefficients that are calibrated using a cross-validation technique. Estimated values are written as a linear combination of neighboring data points and the variance is calculated with the help of the variogram. Such interpolated values are unaffected by the variance quantification procedure.

We then use the PLI model to calculate the variance of a yearly nutrient load under the assumption that only the nutrient

[1] Resident Consulting Scientist, Dept. of Water Resources Evaluation,
South Florida Water Management District, 3301 Gun Club Road.
P.O. Box 24680, West Palm Beach, FL. 33416-4680

[2] Member, ASCE. Senior Environmental Scientist, Dept. of Regulation,
South Florida Water Management District, 3301 Gun Club Road.
P.O. Box 24680, West Palm Beach, FL. 33416-4680

[3] Senior Environmental Scientist, Everglades Restoration Dept.,
South Florida Water Management District, 3301 Gun Club Road.
P.O. Box 24680, West Palm Beach, FL. 33416 4680 and,

[4] Lead Environmental Scientist, Dept. of Planning,
South Florida Water Management District, 3301 Gun Club Road.
P.O. Box 24680, West Palm Beach, FL. 33416-4680

concentrations contained missing values. When these results were compared with those from an arithmetic-mean, a flow-weighted mean, and a linear regression model, the PLI model was found to be comparable to the other three models in terms of variance.

Selection of an appropriate model depends on the characteristics of the data set. Knowing the variance of estimated loads can help regulatory agencies make better decisions to determine whether water quality in the environment is in compliance with established standards or criteria.

INTRODUCTION

Development of rules and regulations to protect environmental resources and monitoring of those resources to evaluate compliance, often require determinations of nutrient loadings. Loading is calculated as the product of discharge and concentration. When concentration and discharge measurements are "complete" and "accurate", loading calculations are straightforward and "accurate". In reality, however, concentration and discharge measurements involve errors and data are often missing. Consequently, all nutrient loadings are approximations of the true loads, with uncertainty resulting from the estimation of missing values. Various models can be used to estimate missing data. To evaluate data that are collected as part of a regulatory program, we believe that all loading models should include some means of quantifying the degree of uncertainty associated with the results.

Many studies have been conducted during the last two decades to improve the accuracy of nutrient loading estimates. Based upon a comparison of some commonly used load calculation methods and their relative errors, Scheider et al. (1978) recommended the use of measured concentrations as the midpoints for estimating phosphorus concentrations in individual time intervals. Cohn et al. (1989) compared the biases and variances of three log-linear regression models and recommended a minimum unbiased load estimator. Stack and Belt (1989) reported that large differences in pollutant loads could result from selection of different flow averaging periods.

Preston et al. (1989) evaluated three broad classes of tributary loading estimation methods: simple averaging, ratio estimation, and regression. They found that none of these estimators was superior to the others for the tested cases. Preston et al. (1992) reported that Beale's ratio estimator was the only method that provided unbiased estimates for both stable and responsive systems, although stratification was necessary under event sampling. By separating precision from accuracy, Rekolainen et al. (1991) concluded that load calculation methods based on summing the products of regularly sampled flows and concentrations resulted in the highest precision, whereas the best accuracy was achieved using methods based on multiplying annual flow by the flow-weighted annual mean concentration. In an attempt to use a less expensive time-composite sampling method for regulatory purposes, Shih et al. (1994) discussed the bias and accuracy of this method and the covariances that occur between discharge volumes and concentrations.

Most of these previous studies have assumed that complete data sets were available (all data necessary to obtain an error-free reference) and used sub-sampling methods to numerically infer their results. In practice, however, complete data sets are rarely available, and various methods, such as piece-wise linear interpolation (PLI), are used to estimate missing data.

PLI is a simple procedure. Two neighboring known data points are connected with a straight line, and estimated values between these known points are read off the line. Due to the deterministic nature of PLI, uncertainties of estimated values are often ignored. In this paper, we develop a model to quantify the uncertainty of values obtained by PLI.

There are two basic approaches for estimating missing data: (1) interpolation, of which PLI is the simplest method, and (2) regression, where linear regression (LR) is the simplest model. Given N data points, PLI defines N-1 straight line equations, each line connecting the two nearest neighboring points. LR defines only one straight line equation with over-determined data points. We focus our model development on PLI, but numerical examples of various models are presented for comparison.

In applying the models for load calculations, we assume that (1) the available flow and concentration data are error-free (data errors are not considered); (2) that only concentration measurements have missing values, so that PLI and other estimation models are applied to concentration data only; and (3) that all concentration data were obtained from (instantaneous) grab samples.

METHODS

Definitions

Let l_i be the load in a given time period, dt; t_i be the midpoint of this period; and C(t) and Q(t) be the concentration graph and discharge hydrograph, respectively. Then,

$$l_i = \int_{t_i-dt/2}^{t_i+dt/2} C(t)\,Q(t)\,dt \tag{1}$$

When dt is sufficiently small, C(t) for the *i*th period can be approximated by a constant C_i. Thus,

$$l_i = C_i \int_{t_j+dt/2}^{t_i-dt/2} Q(t)\,dt = C_i V_i \tag{2}$$

where V_i is the integrated discharge volume in the *i*th time interval. If the time interval, dt, is one day, then C_i and V_i in Equation (2) can be viewed as their daily averages and l_i is the daily load for the day t_i. For a given time period the total load, L, can be obtained by summing the daily loads as follows:

$$L = \sum_{i=1}^{N} C_i V_i \tag{3}$$

where N is the total number of days in the given time period. When daily C_i and V_i data are "complete" and "accurate" for the period, the summation in Equation (3) is straightforward. "Complete" means that there are no missing data for the entire period, whereas "accurate" means that the data are unique and are not subject to any sampling or analytical errors. If either of the these conditions is not met, Equation (3) will incur uncertainty.

Concentration and flow measurements are usually acquired independently. For an accurate but incomplete data set with missing values, uncertainty in the load calculation only occurs due to the estimation of missing values. If n is the number of missing data and K is the number of days when concentration data are available, then Equation (3) can be rewritten as:

$$L = \sum_{i=1}^{n} [C_i]\, V_i + \sum_{k=1}^{K} C_k V_k \qquad (4)$$

where $[C_i]$ is the estimated value of the missing concentration for day *i*, and n + K is the total number of days in the calculation period. The first term on the right-hand side of Equation (4) includes all estimated concentrations while the second term includes only known values. Therefore, the uncertainty of L only comes from the first term.

Models

Before developing the PLI model, some other established models that are commonly used in nutrient load calculations should be examined. Arithmetic mean (AM), flow-weighted mean (FM) and linear regression (LR) models all have established uncertainty analysis procedures.

A. Models that do not fill in missing data

If missing and available data are from the same population, the mean and the variance of this population can be estimated from the available data without estimating individual missing values. Thus, the load for a given period is the sum of the products of the daily discharge volume and the mean concentration during that period. AM and FM models are in this category. These means and variances can be calculated by following procedures in Snedecor and Cochran (1980, pp27-30). The difference between AM and FM is that, for AM, the frequency of each sample is one; for FM, the frequency is proportional to the flow volume at the time that the sample is taken. Since we are considering that only the concentration (SV) has uncertainty, these models are denoted as SVAM and SVFM.

B. Models that fill in missing data

Although the SVAM model and the SVFM model are easy to use and often give satisfactory results, differences between the sample and population means can be significant, especially when the sample size is small. On the other hand, nutrient concentration is often related to other hydrologic parameters such as discharge volume and rainfall. These relationships, which are not used in the AM and FM models, can be used to provide better estimates of missing concentrations and to reduce uncertainties. Linear Regression (LR) is one of the simplest of such models. A simple regression model can be established using daily data, such as discharges, as the independent variable. With a complete set of daily discharge data, each missing daily concentration can then be estimated by the regression equation. Draper and Smith (1981, pp86-87) provide a clear treatment of how to establish such a model. To find the variance of the load estimated as the product of a flow-volume and the concentration, however, one must consider the variance-covariance of multiplying two correlated random variables. We followed Kendall and Stuart (1977, pp261-262) to calculate the variance of the estimated load denoted herein as the SVLR (single-variable linear regression) model.

The PLI model, when established, also belongs to this category. One can use time or distance as the independent variable in PLI, to reduce the covariance problem in further mathematical operations.

Piece-wise linear interpolation (PLI) model

Due to its simplicity, linear interpolation from two adjacent known points is widely used to estimate missing data. Since it is defined by a deterministic equation, uncertainty is often ignored. To our knowledge, no other procedures have been established to quantify the variance of PLI. Without such a variance quantification procedure, PLI is simply a protocol and is not comparable to other models that provide quantitative variance analysis. Because some regulatory agencies have adopted PLI as their standard in nutrient load calculations, it is imperative that a variance analysis procedure be developed for this method.

Given a set of data (y_k, t_k) for k=1, 2, ..., K, where y_k is an independently measured value of a random variable Y at some abscissa t_k such as time, a missing value z_i at t_i where $t_k < t_i < t_{k+1}$ may be estimated based on the two adjacent known points y_k and y_{k+1} using the following linear interpolation equation:

$$z_i = y_k + (y_{k+1} - y_k) \frac{(t_i - t_k)}{(t_{k+1} - t_k)} \tag{5}$$

The variable Y is assumed to be stationary for the first two moments. It can be seen that when $t_i = t_k$, then $z_i = y_k$; similarly, when $t_i = t_{k+1}$, then $z_i = y_{k+1}$. Our question is, what is the uncertainty or variance of the estimated z_i? This uncertainty results from the inability of Equation (5) to predict the true value of z_i, not from the measurement errors of y_k and y_{k+1}. Equation (5) can be rewritten into a linear combination form as follows:

$$z_i = W_{1,i}\, y_k + W_{2,i}\, y_{k+1} \tag{6}$$

$$W_{2,i} = \frac{t_i - t_k}{t_{k+1} - t_k} - \frac{\Delta_1}{\Delta_1 + \Delta_2} \tag{7}$$

$$W_{1,i} = \frac{t_{k+1} - t_i}{t_{k+1} - t_k} = \frac{\Delta_2}{\Delta_1 + \Delta_2} \tag{8}$$

where $W_{1,i}$ and $W_{2,i}$ are the weight factors of y_k and y_{k+1}, respectively. It is clear from equations (7) and (8) that $W_{1,i}$ and $W_{2,i}$ satisfy the following constraint of unbiasedness:

$$W_{1,i} + W_{2,i} = 1. \qquad (9)$$

Therefore, the weights are determined when the t_i is known. Equation (6) can be expanded to include linear combinations of three or more known data points.

To evaluate the variance of z_i and thus the variance of the calculated load, we have used a modified version of the kriging process as follows:

(1) An appropriate theoretical semi-variogram (Skrivan and Karlinger, 1980) is selected which is a function of the variable Δ.

(2) The selected theoretical semi-variogram is calibrated by cross-validation (Hjorth, 1994, pp 24-57) under the constraint that the variance produced with the semi-variogram agrees with the variance produced by linear interpolation of the K known points. In this step, a two-point kriging procedure (Journel, 1989) is followed to determine the variances of the estimates. Steps (1) and (2) are repeated until a satisfactory semi-variogram is obtained.

(3) Variance of the load is calculated using the calibrated theoretical semi-variogram.

For example, we can select an exponential function as the theoretical semi-variogram in step (1):

$$\begin{aligned} \gamma(\Delta) &= 0\,, & \Delta &= 0 \\ \gamma(\Delta) &= \omega(1 - e^{\Delta/\alpha}) + c\,, & \Delta &> 0 \end{aligned} \tag{10}$$

where ω, α, and c are coefficients to be calibrated. The maximum of the above function is the sill (ω+c, when Δ → ∞). The parameter c is called the "nugget effect" that creates a jump of uncertainty if not exactly at the measured data point. Step (2) is to calibrate the three coefficients by cross-validation. In this process, the inner (K-2) known points, y_k (k=2, 3, ..., K-1), are first estimated one at a time using Equation (6) as follows:

$$\hat{z}_k = W_{1,k}y_{k-1} + W_{2,k}y_{k+1} \tag{11}$$

where $\hat{z}_k$ is the estimate of y_k. Non-neighboring points can also be used as long as their weights are correctly calculated. In most cases, however, it is desirable that the selected semi-variogram be valid for small Δ's. Therefore, it is advantageous to use the nearest points in Equation (11). The sum of the squared cross-validation error is calculated using the following equation:

$$S_{\hat{z}} = \sum_{k=2}^{K-1} (y_k - \hat{z}_k)^2 \tag{12}$$

The variance is:

$$\sigma_{\hat{z}}^2 = \frac{S_{\hat{z}}}{K-2} \tag{13}$$

Equation (13) is comparable to variances of other models when the variances are defined as the average of the sum of squared differences between observed and predicted errors. Equations (12) and (13) are immediately known for a given data set after the weights are defined. The sum of squared errors in Equation (12) is used as the constraint to calibrate the coefficients in the theoretical semi-variogram.

Let $\gamma(\omega, \alpha, c; \Delta)$ be the semi-variogram in Equation (10). The constraint in step (2) requires that the sum of variance produced with the theoretical semi-variogram agrees with equation (12). We now express the residual of the estimate $\hat{z}_k$ in terms of the semi-variogram. From Equation (11):

$$e^2_{\hat{z}_k} = \{ y_k - [W_{1,k}y_{k-1}+W_{2,k}y_{k+1}] \}^2 \tag{14}$$

Expanding Equation (14) and considering $W_{1,k} + W_{2,k} = 1$,

$$e^2_{\hat{z}_k}=2W_{1,k}\frac{(y_k-y_{k-1})^2}{2}+2W_{2,k}\frac{(y_k-y_{k+1})^2}{2}-2W_{1,k}W_{2,k}\frac{(y_{k-1}-y_{k+1})^2}{2} \tag{15}$$

Based on the semi-variogram,

$$\gamma(\Delta) = \frac{1}{2}Var[(y_k - y_{k+\Delta}] = \frac{1}{2}(y_k - y_{k+\Delta})^2 \tag{16}$$

Therefore, the expected values of the squared differences in Equation (15) can be approximated by the theoretical semi-variogram $\gamma(\Delta)$ as the following:

$$c^2_{\hat{z}_k} = 2W_{1,k}\gamma(\Delta_{1,k}) + 2W_{2,k}\gamma(\Delta_{2,k}) - 2W_{1,k}W_{2,k}\gamma(\Delta_{1,k}+\Delta_{2,k}) \tag{17}$$

Equation (17) is the variance produced with the semi-variogram. The imposed constraint becomes:

$$\sum_{k=2}^{K-1} e_{z_k}^2 = S_z \tag{18}$$

The coefficients, ω, α, and c, for the theoretical semi-variogram $\gamma(\omega,\alpha,c;\Delta)$ are calibrated to satisfy Equation (18). A two-point kriging procedure is followed to derive a set of raw semi-variance data at selected intervals from a given data set. The theoretical semi-variogram is fitted to this data set by least squares to obtain the initial values of the coefficients, ω_0, α_0, and c_0. The initial values are then adjusted so that Equation (18) is satisfied. Another constraint used in the cross-validation is that the sill is kept constant during the coefficient adjustment, i.e., $\omega+c = \omega_0+c_0$. Keeping the sill constant for a monitoring site ensures that the variance of a calculated load is only dependent on the number and sequence of missing data points.

Once the coefficients are determined for the theoretical semi-variogram, proceed to step (3) to calculate the variance of the estimated load. An equation similar to Equation (17) is used, except that the subscript k (indicating known points) is replaced by i (indicating missing points) and the equation is multiplied by a factor equal to the squared discharge volume:

$$Var(l_j) = 2V_i^2\ [W_{1,i}\gamma(\Delta_{1,i})+W_{2,i}\gamma(\Delta_{2,i})-W_{1,i}W_{2,i}\gamma(\Delta_{1,i}+\Delta_{2,i})] \tag{19}$$

If the time interval, t_i, is one day, then the above variance is the variance of the calculated daily load. The variance of the calculated yearly load then is expanded as:

$$Var(L) = \sum_{i=1}^{n} Var(l_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} V_i V_j \; [(\omega+c)-\gamma(\Delta_{i,j})] \;\; ; \;\; i \neq j \quad (20)$$

where n is the number of missing concentrations. The upper bound of the variance in Equation (20) is the sill ω+c, and $\Delta_{i,j}$ is the time difference in days between day i and day j. The second term on the right-hand side of Equation (20) is the sum of the covariances between estimated concentrations, i.e., $Cov(C_i, C_j) = (\omega + c) - \gamma(\Delta_{ij})$. Equations (15) through (20) define a PLI model that is denoted as the Single Variance Linear Interpolation (SVLI) model.

RESULTS AND DISCUSSION

The models discussed above were applied to a sample data set from water control structures located near the southern end of Florida over a 13-year period (1978-1990). Characteristics of this data set are described in Table I. The data set has complete daily discharge (no missing data) but incomplete daily concentration data. In the SVLR model, a simple equation ($C = a_0 + a_1V$) using daily discharge as the independent variable, was used to estimate missing daily concentrations. In the SVLI model, the exponential function in Equation (10) was used as the semi-variogram.

Yearly loads were also calculated using each model. To compare the load variances, we also defined and calculated the following dimensionless coefficient of load variation,

$$G = \frac{1}{L}\sqrt{Var(L)} \qquad (21)$$

where Var(L) is the load variance and L is the calculated load. Results are plotted in Figures 1, 2 and 3.

Differences among the loads calculated by these models are small in magnitude (Figure 1). In 1982, many high concentration values were recorded on low discharge days. Therefore, the SVAM model gave a 60% higher estimate than the other three models in that year.

The G values from different models are significantly different (Figure 2). G values from the SVAM and SVLR models were much lower than those from the SVFM and SVLI models. The G values of a single model also varied annually. One reason for this variation is that the uncertainty of a model is data dependent, i.e., differences in data values and in the number of missing data will result in different G values for different years. Another reason is that the uncertainties of missing concentrations propagate differently in different models into the final load variance, Var(L). The SVLI and SVLR models tend to magnify the variance of missing data estimates, while the variance of SVAM and SVFM models depends more on the variation of collected data.

The G values from the different models showed a similar pattern of change over time (Figure 2). This was because all of these models used the same data, so that the number of data available in each year was the same. We expect a good load model to be data-driven.

To understand how different models behave, a numeric index is needed to compare the uncertainties attributable to each model. Uncertainty of missing data estimates can be represented by the coefficient of variation (CV) of model predictions, which is defined as follows:

$$CV = \frac{1}{C_m}\sqrt{\frac{1}{K}\sum_{i=1}^{K} (C_i - C_{p,i})^2} \tag{22}$$

where C_i is the observed concentrations, $C_{p,i}$ is the model predicted value of C_i, C_m is the arithmetic mean of C_i, and K is the total number of observations. The error magnification due to the multiplication by flow-volume can be specified as the difference between G and CV. For this purpose, the DG coefficient is defined as follows to indicate model behavior:

$$DG = \frac{1}{M}\sum_{i=1}^{M} (G_i - CV_i)^2 \tag{23}$$

where M is the number of years (13 in this study). Calculated CV values for the three models are shown in Fig. 3. Comparing Figures 2 and 3, it is clear that a small variance in the missing data estimates does not necessarily correspond to a small variance in the calculated load. Variance of load calculations depends upon model-error and multiplication pattern. DG, as an overall model performance indicator, is shown in Table 2.

The SVLI gives the smallest DG value, which means that the variance of the load by this model varies the least from the variance attributable to data source. Therefore, PLI is the preferred approach. DG is a useful index for model comparisons, and is perhaps also applicable to two-variance cases when both flow and concentration measurements contain missing data.

In the SVLI model, interpolation weights are determined prior to the calculation of the load variances. Other linear interpolation methods with fixed weights have been used by investigators (Scheider et al., 1978). For the purpose of comparison, the following three commonly used fixed-weight linear interpolation methods were also investigated.

a) The mid-point method uses observed data as the midpoints of corresponding intervals to interpolate missing concentrations (equivalent to assigning $W_1 = 1$ and $W_2 = 0$).

b) The equal-weights method interpolates missing concentrations by the arithmetic mean of the two adjacent known points (equivalent to assigning $W_1 = W_2 = 0.5$).

c) The three-points method interpolates missing concentrations by the three nearest neighboring points with fixed weights $W_1 = 0.25$, $W_2 = 0.5$, and $W_3 = 0.25$.

Using the same data set, it was found that these three fixed-weight linear interpolation methods gave very similar yearly load estimates to these provided by the SVLI model, but had higher G and DG values. DG valued for the three-points method was 0.052, for the mid-point method was 0.169.

SUMMARY AND CONCLUSIONS

In this paper, we propose a model to quantify the variance in piece-wise linear interpolation (PLI) and apply the model to nutrient load calculations. This model provides a method for regulatory agencies that use PLI to compare their results with those from other models. The PLI uncertainty analysis procedure consists of the following steps:

a. Estimate missing data using PLI with Equation (6). The load can be determined before the variance analysis is conducted.

b. Select a theoretical semi-variogram based on properties of the data and calibrate it for each year by cross-validation until the convergence requirement in Equation (18) is met. This convergence requirement ensures that the variances derived from the model are comparable to those obtained from other models.

c. Compute variance of the calculated yearly load using equation (20).

By comparing the SVLI model with other models, we concluded:

1. Load estimates derived by applying different models to the sample data set were not significantly different.

2. Uncertainties in missing data estimations were dependent upon the type of model used and data properties.

3. The dimensionless coefficient of load variation, G, in Equation (21) was a convenient way to compare the uncertainties in loading estimations among different models, especially when both discharge and concentration measurements contained missing values. This coefficient also provides information on error magnification,

resulting from error propagation within a particular model.

4. The DG coefficient defined in Equation (23) can be used to select a numerically robust model. A smaller DG suggests less magnification of the uncertainty in the missing data estimation. Consequently, a perturbation in data will not significantly change the loading estimation result of a model if its DG value is small. The SVLI model gave the smallest DG value and thus was the most desirable model for the data set that was used in this study.

5. Other fixed-weight linear interpolation models were comparable to the SVLI model in terms of calculated loads and load variances. The three-points model, in particular, provided relatively smoother transitions from point to point.

6. Since the models are data dependent, there is no guarantee that the best model for one data set will also be best for another set. Model selection depends on how well the model describes the inherent properties of the data.

One may argue that since calculated loads from different models are often very close to each other, it doesn't matter which model is used. Statistically this is true. However, in the context of a regulatory program, specific numeric limits are often set as loading thresholds. In such cases, knowing the uncertainty or confidence interval of an estimated load may be critical to regulatory agencies in their determination of whether a loading estimate is, or is not, in compliance with established criteria.

APPENDIX I: REFERENCES

Cohn, T. A., DeLong, L. L., Gilroy, E. J., Hirsch, R. M, and Wells, D. K. (1989). "Estimating Constituent Loads." Water Resources Research, 25(5), 937-942.

Draper, N. R., and Smith, H. (1981). *Applied Regression Analysis.* 2nd ed., John Wiley & Sons, Inc. New York.

Hjorth, J.S. Urban (1994). *Computer Intensive Statistical Methods.* Chapman and Hall. New York.

Journel, A. G. (1989). "Fundamentals of Geostatistics in Five Lessons." American Geophysical Union, Washington, D.C.

Kendall, M. and Stuart, A. (1977). *The Advanced Theory of Statistics*. Volume 1: Distribution Theory. 4th Ed., MacMillan Publishing Co., Inc. New York.

Preston, S. D., Bierman, V. J., and Silliman, S. E. (1989). "An Evaluation of Methods for Estimation of Tributary Mass Loads." Water Resources Research, 25(6), 1379 - 1389.

Preston, S. D., Bierman, V. J., and Silliman, S. E. (1992). "Impact of Flow Variability on Error in Estimation of Tributary Mass Loads." J. Environ. Engrg. ASCE, 118(3), May/June, 402-419.

Rekolainen, S., Maximilian P., Kamari, J., and Ekholm, P. (1991). "Evaluation of the Accuracy and Precision of Annual Phosphorus Load Estimates from two Agricultural Basins in Finland." J. of Hydrology, Vol. 128, 237-255. Kluwer Academic. Netherlands.

Scheider, W. A., Moss, J. J., and Dillon, P. J. (1978). "Measurement and Uses of Hydraulic and Nutrient Budgets. In: Lake Restoration." Proceedings of a National Conference, August 22-24, 1978. Minneapolis, Minnesota.

Shih, G., Abtew, W., and Obeysekera, J. (1994). "Accuracy of Nutrient Load Calculation Using Time-composite Sampling." Transactions of the ASAE, March/April, Vol. 37 No.2, pp419-429.

Skrivan, J. A. and Karlinger, M. R. 1980. *Semi-variogram Estimation and Universal Kriging Program.* U.S.G.S. Water Resources Division, Tacoma, Washington.

Stack, W. P., and Belt, K. T. (1989). "The Selection of Appropriate Flow Averaging Periods in Evaluating Pollutant Loading Using the Flow Interval Method." Lake Reservoir Management, 5(2),67-73. Madison, Wisconsin.

Snedecor, G. W, and W. G. Cochran (1980). Statistical Methods. The Iowa State University Press. Ames, Iowa.

APPENDIX II. NOTATION

The following symbols are used in this paper.

C	concentration
c	coefficient of semi-variogram
$\overline{C}_m$	yearly mean concentration
Cov	co-variance
CV	coefficient of variation
DG	coefficient of deviation of G from CV
dt	time increment
G	Coefficient of load variation, dimensionless
K	total number of known data points
L	yearly load
l	daily load
l_m	mean daily load
N	total number of days of calculation period
n	total number of missing data points
Q	discharge
S	sum of squared errors
t	time
V	discharge volume
Var	variance
V_m	mean daily discharge volume
W	weight of interpolating point
y_k	predicting or known point in linear interpolation
$\hat{y}_i$	predicted value of a variable by linear regression method
z_i	predicted value of a missing data by linear interpolation

$\hat{z}_k$	predicted value of a known data point
[.]	an estimated value, involving variance
α	coefficient of semi-variogram
Δ	difference in sampling dates of two samples, normally in days
γ	semi-variogram
σ^2	variance
ω	coefficient of semi-variogram

TABLES

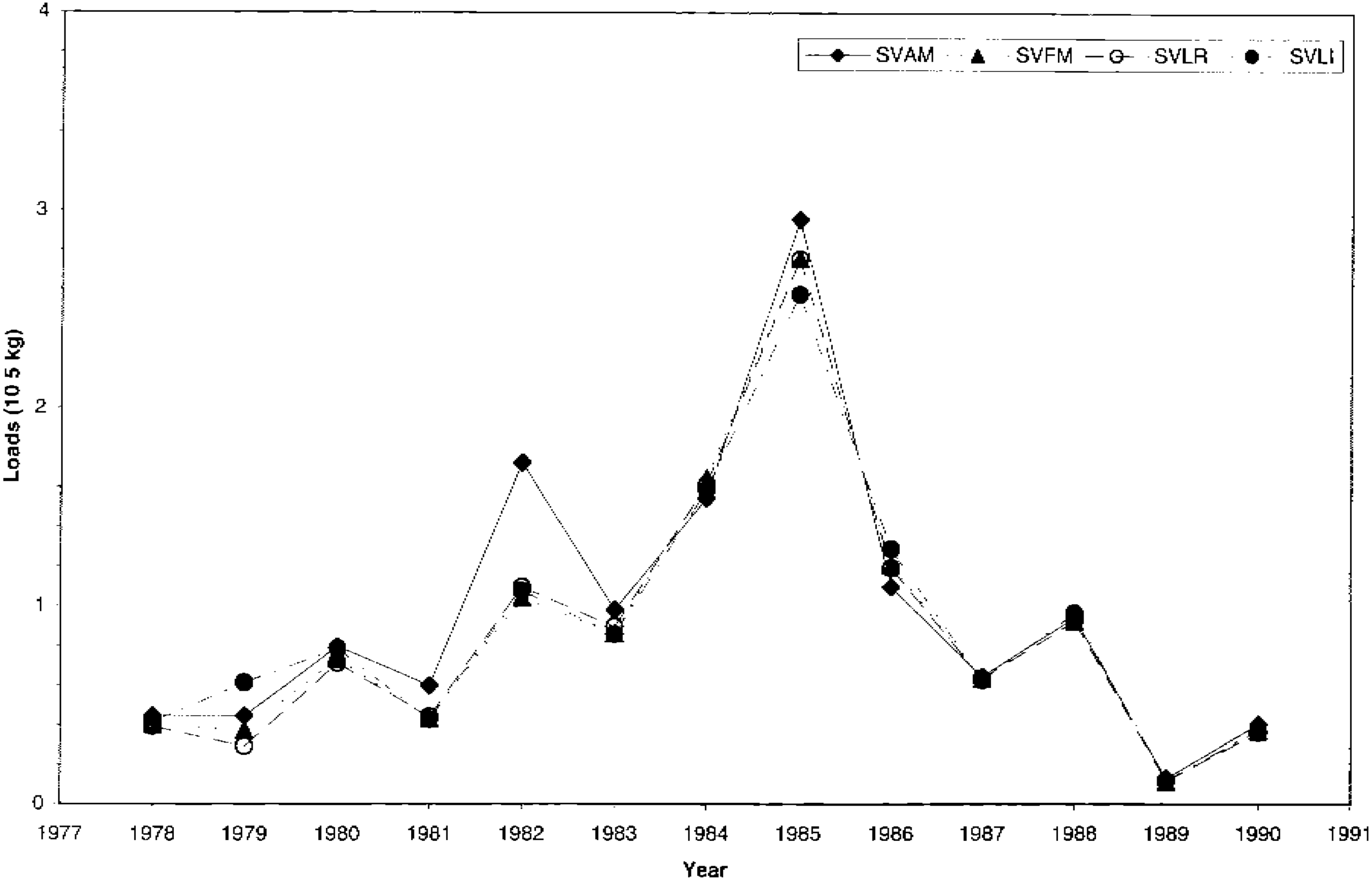
FIGURES

TABLE 1. A Description of Data Used in the Example

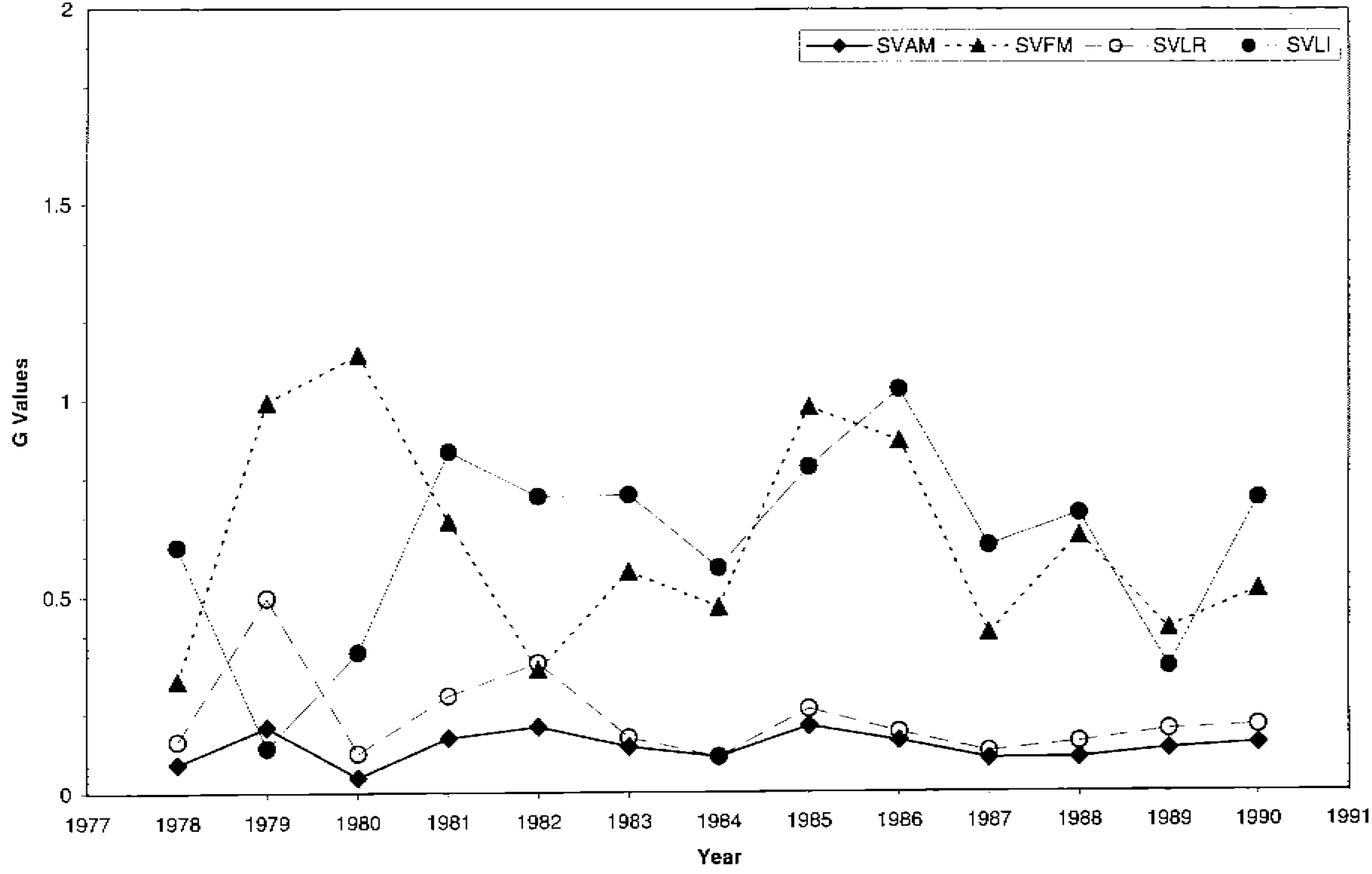
Year	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Water volume (10^9m^3)	.691	.526	.730	.353	1.40	1.10	1.01	.719	.659	.474	.607	.084	.216
No-flow days	0	0	33	0	0	0	0	23	0	91	48	265	152
Max. Flow(10^6m^3/day)	7.00	6.26	6.93	2.99	10.9	8.60	5.56	5.41	6.74	3.68	3.68	1.73	2.42
Min. Flow(10^6m^3/day)	.093	.149	0	.051	.031	.947	.408	0	.168	0	0	0	0
Number of samples	26	65	80	23	21	24	25	20	25	18	27	8	14
Serial-sample days	0	31	61	0	0	0	0	0	0	0	0	0	0
Max. [P] (ppb)	150	280	364	459	400	286	285	970	558	286	286	215	390
Min. [P] (ppb)	10	10	40	80	47	40	44	48	40	79	83	84	93

Table 2. Model Performance Indicator, DG.

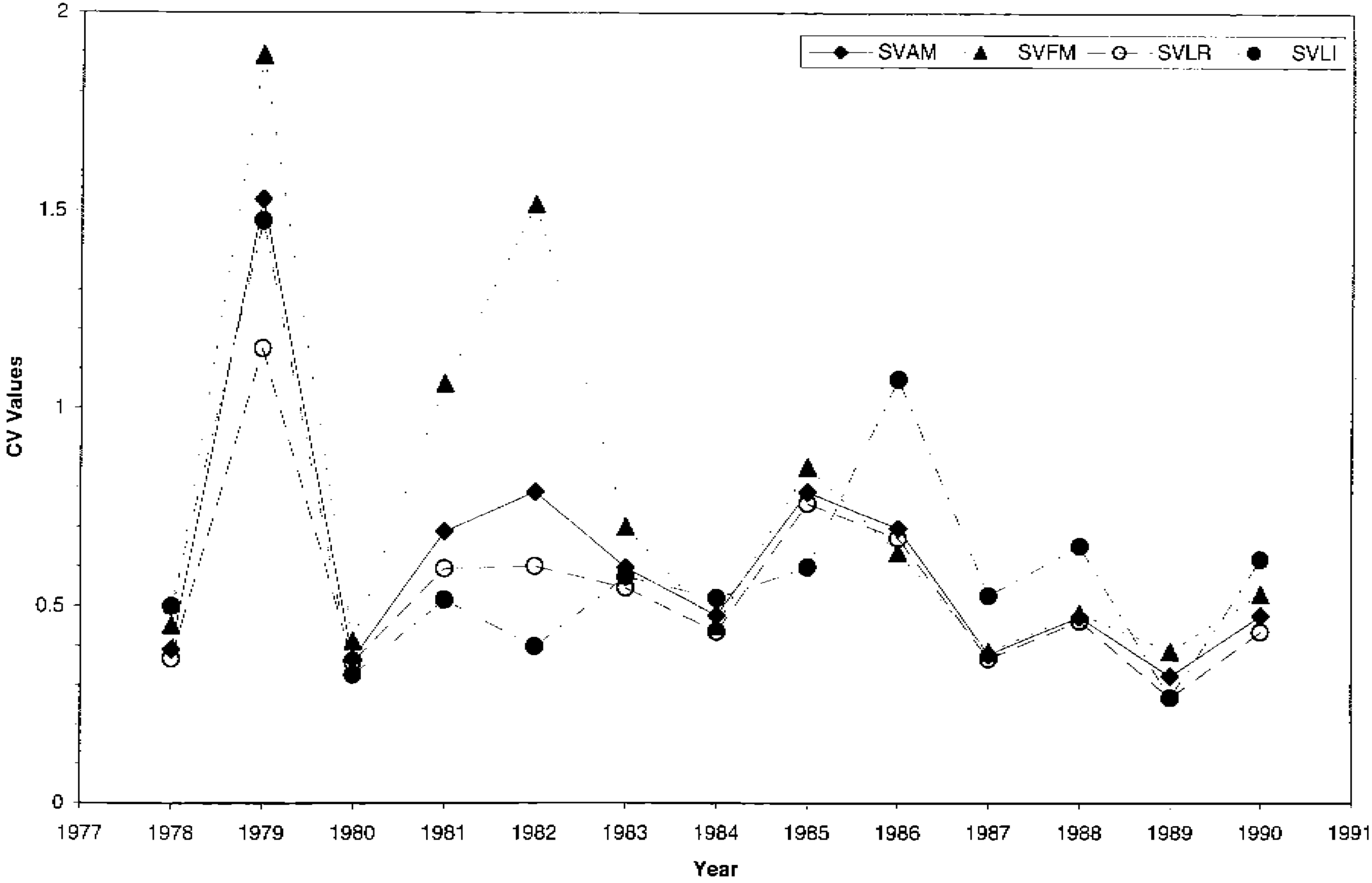
Model:	DG
SVLI: single-variable Linear Interpolation	0.040
SVLR: single-varibale Linear Regression	0.189
SVFM: single-variable Flow-weighted Mean	0.238
SVAM: single-variable Arithmetic Mean	0.328


SVAM
SVFM
SVLR
SVLI
Loads (10 5 kg)
0
1
2
3
4
Year
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991

[illegible] Fig 2


SVAM
SVFM
SVLR
SVLI
G Values
2
1.5
1
0.5
0
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
Year

Shin, Fig 3



◆ SVAM
▲ SVFM
○ SVLR
● SVLI
CV Values
2
1.5
1
0.5
0
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
Year